NAMED ENTITY RECOGNITION ON TRANSCRIPTION USING CASCADED CLASSIFIERS

Firoj Alam[1,2]

[1]FBK-irst,
via Sommarive 18, I-38123 Povo (TN), Italy
[2]SIS Lab, Department of Information Engineering and Computer Science,
University of Trento, 38123 Povo (TN), Italy
firojalam@gmail.com

Abstract. This paper presents a Named Entity Recognition (NER) system on broadcast news transcription which is a combination of two different classifiers. In addition, we present a comparative analysis of the results obtained by extracting Named Entities from two different types of documents: written documents and spoken documents. Written documents are documents in which text appears as standard written form e.g. newspaper articles. Spoken (transcribed) documents are the documents where orthographic information and punctuation are missing. In transcribed documents, an absence of these two main features often causes a drop in performances to recognize Named Entities (NEs). An additional error in the transcription made by the Automatic Speech Recognition (ASR) system is that it is not able to recognize the right sequence of words. This also introduces additional performance reduction of NER. The system performed the best on the task of Italian NER at Evalita 2011 with F1 of 63.50%. Obtained results of this study are going to be considered for integration into Typhoon [3], a NER system developed by HTL group at FBK, to deal with transcribed broadcast news too.

Keywords: Named Entity Recognition, Entity Detection, NER

1 Introduction

In this study we report NER on broadcast news transcriptions. The first step in most IE (Information Extraction) tasks is to detect and classify all the proper names mentioned in a text. Named entity means anything that can be referred to a proper name; typically names of person, organization, location, time, date and money. This process of named entity recognition refers to the combined task of finding spans of text that constitute proper names and then classifying the entities being referred to according to their type [1]. The NER system takes an unlabeled text as input and produces an annotated block of text recognizing entities.

The goal of this work is to study the NER on transcription in accordance with the study of an existing NER system, named as "Typhoon" to identify the named entities from written texts and from spoken texts (broadcast news transcription). Written text means the text which appears in standard written form e.g. newspaper articles. Whereas, the spoken text means - the text that comes as transcription from automatic speech recognition [2]. The transcribed texts lack orthographic information – documents that are lower case or upper case. It also lacks punctuation marks. Orthographic, case and punctuation marks are the key features of NE tagging. The absence of these features in spoken document (e.g. transcription) typically results in poor performance [2]. There is an additional error in transcription made by the ASR system in which it is not able to recognize the right sequence of words. This also introduces additional performance reduction of NER.

In this study, we combined two different algorithms in cascade, such algorithms are Support Vector Machine (SVM) and Conditional Random Field (CRF) [2], those are the state of art and the most widely used algorithms in NER task. We also used CRF to exploit unlabelled data and for case and punctuation restoration.

The rest of the paper is organized as follows. Section 2 discusses task description followed by the architecture of the system in section 3. Section 4 explains the experiments that we carried out. In section 5, we present the results. Finally, discussion and conclusion appear in section 6.

2 The Task Description

In the Named Entity Recognition (NER) task at Evalita 2011, systems are required to recognize different types of Named Entities (NEs) in Italian broadcast transcribed news texts. Such types are Person (PER), Organization (ORG), Location (LOC) and Geo-Political Entities (GPE). Training and test data consist of spoken news of about five hours of transmission, for a total of about 40,000 words of each that are recorded and transcribed [8]. The training and test data of spoken news were provided by the local broadcaster RTTR[1]. In addition, training data also consists I-CAB, i.e. the corpus of (written) news stories taken from local newspaper L'Adige and annotated with Named Entities used for the NER tasks at Evalita 2007 and Evalita 2009. The I-CAB corpus contains 314,260 tokens. The format of the I-CAB corpus consists of token, POS, file id and NE tag; and the format of the transcription consists of token, file id and NE tag. The use of external resources was allowed for open task and systems were evaluated in terms of precision, recall and F1 measure by using the CoNLL 2002 scorer.

3 Architecture of the System

We have followed the same approach as Typhoon [3], a system for NER developed by HLT unit at FBK, which uses two classifiers in cascade to exploit data redundancy and patterns extracted from a large text corpus. In the first step, the system uses

[1] RTTR - http://www.rttr.it/

CRF++[2], which is an implementation of Conditional Random Fields [5] to recognize NEs in a large and unlabeled corpus. In the second step, SRILM *disambig*[3], implementing Hidden Markov Models (HMM), exploits the annotation made by the first classifier to annotate NEs both in the training set and in the corpus to be annotated. Finally, CRF++ uses the data in second phase as features and performs annotation. However, in this study our attempt is to overcome the bottleneck of *disambig* in order to improve the performance of Typhoon. We want the annotation tag to depend not only on the previous tag but also on the context (i.e. words preceding and following the word to be annotated). For this reason, we tried to use CRF++ instead of *disambig* as second classifier whereas Yamcha[4] was tested as an alternative to CRF++ for implementing the first classifier. Moreover, we used CRF++ for case and punctuation restoration as well. The architecture of this system is given in **Fig. 1**.

The features we used in the experiments are token, lower case form of the token, lemma, POS, case info, second classifier feature and gazetteers. In order to extract POS and lemma we used TagPro [6]. Gazetteers are extracted from Italian phonebook, Wikipedia, various web sites of Italy and Trentino, Italian and American stock market and geographical location from Wikipedia. We use large and unlabeled corpus to add an additional feature. Additionally, we implemented case and punctuation restoration model.

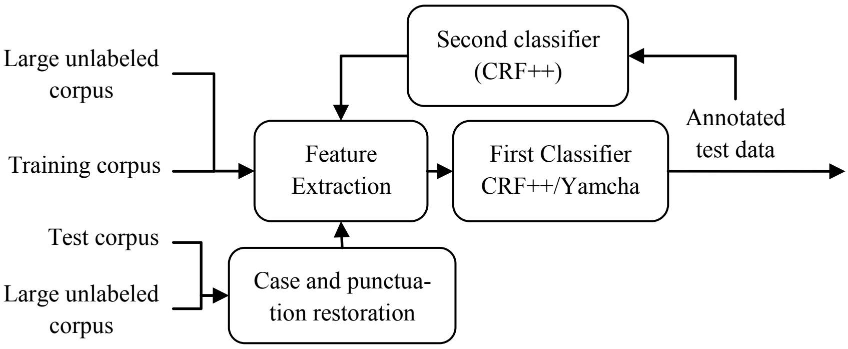

Fig. 1. Architecture of the system

4 Experiments

4.1 Transcription vs Newspaper

Recognizing entities from written text (e.g. newspaper articles) is fairly straightforward, since it has all the relevant information to recognize entities. However, in transcription, we don’t have all the required information to recognize named entities. The

[2] CRF++ - http://crfpp.sourceforge.net
[3] http://www.speech.sri.com/projects/srilm
[4] http://chasen.org/~taku/software/yamcha/

transcription (ASR output) greatly suffers due to out-of-vocabulary words, insertion of erroneous words and speech disfluencies [4]. Speech disfluencies are hesitations, field pauses, insertion of false starts which basically reduce the quality of the transcript and in turns reduce the quality of NER. We have tried to see the difference between newspaper and transcription of NEs recognition, where our training data is based on newspaper text.

In this experiment, training model has been generated using Evalita-2007 data set and Evalita-2009 training set. Then, the system is tested on Evalita-2009 test set (newspaper text), manually and automatically transcribed broadcast news (Evalita 2011 data set). It is observed that the performance of the system is better for newspaper content compare to speech transcription as shown in Table 1.

There are considerable differences between manual and automatic transcription. In automatic transcription we don't have case and punctuation information as much as manual transcription has. So we need to preprocess the data before extracting the features. This preprocessing involves case-restoration, removing or adding punctuation and so on.

It is observed that there are some ASR errors in transcription which leads the system to recognize NEs incorrectly. ASR makes three types of errors (edit operations) such as insertion (I), deletion (D) and substitution (S) [8]. The word error rate (WER) of the ASR on test set (automatically transcribed) is 16.39%, unit accuracy is 83.61% and percent correct is 87.48%. The evaluation script of the ASR converts both gold standard and system transcription into lower case form and then uses Levenshtein distance to align tokens. After that, it performs evaluation [8] using CONLL scorer. As it ignores case information so WER is lower, whereas case information is important in NER.

Table 1. Performance of the system on newspaper texts and transcription

Data	Precision	Recall	F1
Newspaper text (Evalita-2009 test set)	85.19%	81.88%	83.50
Manual transcription	80.06%	75.52%	77.72
Automatic transcription	63.50%	58.05%	60.66

4.2 Case and Punctuation Restoration

We implemented a case and punctuation restoration model using a corpus of 250 millions tokens taken from the newspaper L'Adige. We used one previous and one following word in context and used CRF as a classifier. The performance of our case and punctuation restoration model is 96.49 (F1) on L'Adige corpus of 10 million tokens. This model is used to restore case and punctuations. Then, I-CAB data is used to train the system and Evalita 2011 test set is used for testing. We achieved an improvement of the system after case restoration as shown in Table 2.

Table 2. Performance of the case restoration system

Automatic transcription	Precision	Recall	F1
Without case restoration	63.84%	58.91%	61.27
With case restoration	66.53%	58.15%	62.06

5 Results

Results have been measured in terms of precision, recall and F-measure as instructed in the guideline [8]. This system has performed best in the NER task of Evalita-2011. Table 3 and Table 4 show the official results on the closed and open tasks on Evalita-2011 test set. Whereas, Table 5 shows the official results of the manually transcribed Evalita-2011 test set. In final ranking, manually transcribed data were not taken into consideration. In open task, we have not used case restoration as we were getting lower performance on the manual transcription due to case restoration. Since in manual transcription, we do have case information and so after case restoration it degrades case information. However, after evaluation campaign we tested our system using case restoration on the test set and we achieved an improvement of F1 64.28.

Table 3. Official results on closed task

Category	Precision	Recall	F1
Overall	61.76%	60.23%	60.98
GPE	81.79%	78.52%	80.12
LOC	65.22%	47.87%	55.21
ORG	50.21%	43.85%	46.82
PER	47.28%	55.26%	50.96

Table 4. Official results on open task

Category	Precision	Recall	F1
Overall	65.55%	61.69%	63.56
GPE	80.33%	80.44%	80.38
LOC	76.36%	44.68%	56.38
ORG	60.51%	47.52%	53.24
PER	48.92%	54.39%	51.51

Table 5. Official results of the manually transcribed Evalita-2011 test set

Task	Precision	Recall	F1
Closed task	79.33%	79.80%	79.57
Open task	82.82%	81.27%	82.04

6 Discussion and Conclusion

In this paper, we described our study of NER on transcription. It is observed that the performance of NER on transcription is worse than newspaper text. However, we used several approaches to improve the performance and the system achieved best results in the evaluation. The second classifier feature and case restoration gives better performance of the system which can be explored further to achieve better results. Moreover, second classifier methodology can also be applied using transcribed unlebeled data which is one direction to look at in future. In addition, we can adapt relevant examples from the test set into the training set which is left for future study.

Acknowledgments. This work has been partially supported by the project Live-Memories, funded by the Provincia Autonoma of Trento. Thanks to Roberto Zanoli, HLT unit, FBK for his collaborative effort.

References

1. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing Computational Linguistics and Speech Recognition (2ed.). Prentice Hall (2008)
2. Rohini, K., Srihari, C.N., Li, W., Ding, J.: A Case Restoration Approach to Named Entity Tagging in Degraded Documents. In: Seventh International Conference on Document Analysis and Recognition (ICDAR'03), vol. 2, pp. 720--724 (2003)
3. Zanoli, R., Pianta, E., Giuliano, C.: Named Entity Recognition through Redundancy Driven Classifiers. In: Proceedings of Evalita 2009, Reggio Emilia (2009)
4. Favre, B., Béchet, F., Nocéra, P.: Robust Named Entity extraction from large spoken archives. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA (2005)
5. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning, pp. 282--289. Morgan Kaufmann, San Francisco, CA (2001)
6. Pianta, E., Zanoli, R.: TagPro: A system for Italian PoS tagging based on SVM. Intelligenza Artificiale, Special Issue on NLP Tools for Italian, vol. IV, issue 2 (2007)
7. Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Bartalesi Lenzi, V., Sprugnoli, R.: I-CAB: the Italian Content Annotation Bank. In: Proceedings of LREC 2006 (2006)
8. Bartalesi Lenzi, V., Speranza, M., Sprugnoli, R.: Named Entity Recognition on Transcribed Broadcast News - Guidelines for Participants, Evalita 2011 (2011)